MATEMÁTICA COMPUTACIONAL

Cap. 4. Resolução Numérica de Sistemas Lineares

Filipe J. Romeiras

Departamento de Matemática
Instituto Superior Técnico

Apontamentos das aulas da disciplina do mesmo nome do 2º ano de
Mestrados Integrados e Licenciaturas em Ciências de Engenharia
do Instituto Superior Técnico

4. RESOLUÇÃO NUMÉRICA DE SISTEMAS LINEARES

Normas matriciais

• Uma vez que $\mathbb{M}^n(\mathbb{R})$ ou $\mathbb{M}^n(\mathbb{C})$ são espaços vectoriais de dimensão n^2, topologicamente equivalentes aos espaço $\mathbb{R}^{n^2}$ ou $\mathbb{C}^{n^2}$, respectivamente, qualquer das normas que introduzimos em espaços vectoriais também serve como norma matricial. No entanto nem todas as normas possíveis têm interesse prático. Vamos, portanto, começar por definir duas condições suplementares que as normas matriciais devem satisfazer para se tornarem úteis.

Definição. Seja M uma norma no espaço $\mathbb{M}^n(\mathbb{C})$. A norma M diz-se **regular** se e só se for satisfeita a condição

$$\|AB\|_M \leq \|A\|_M \|B\|_M, \qquad \forall A, B \in \mathbb{M}^n(\mathbb{C}).$$

Definição. Seja M uma norma no espaço $\mathbb{M}^n(\mathbb{C})$ e V uma norma no espaço vectorial $\mathbb{C}^n$. A norma M diz-se **compatível** com a norma V se e só se for satisfeita a condição

$$\|Ax\|_V \leq \|A\|_M \|x\|_V, \qquad \forall A \in \mathbb{M}^n(\mathbb{C}), \quad \forall x \in \mathbb{C}^n.$$

• A definição seguinte permite-nos obter normas matriciais que satisfazem às cinco condições impostas.

Definição. Diz-se que uma norma matricial M em $\mathbb{M}^n(\mathbb{C})$ está **associada** a uma certa norma vectorial V em $\mathbb{C}^n$ se e só se ela for definida pela igualdade

$$\|A\|_M = \sup_{x \in \mathbb{C}^n \setminus \{0\}} \frac{\|Ax\|_V}{\|x\|_V}. \qquad (\#)$$

Também se diz neste caso que a norma M é a norma **induzida** em $\mathbb{M}^n(\mathbb{C})$ pela norma vectorial V em $\mathbb{C}^n$ ou é a norma em $\mathbb{M}^n(\mathbb{C})$ **subordinada** à norma vectorial V em $\mathbb{C}^n$.

Proposição. A equação (#) define uma norma matricial, regular e compatível com a norma V, isto é, que satisfaz às cinco condições:

(1) $\|A\|_M \geq 0, \quad \forall A \in \mathbb{M}^n(\mathbb{C}); \quad \|A\|_M = 0 \;\Leftrightarrow\; A = 0;$

(2) $\|\alpha A\|_M = |\alpha| \|A\|_M, \quad \forall A \in \mathbb{M}^n(\mathbb{C}), \;\; \forall \alpha \in \mathbb{C};$

(3) $\|A + B\|_M \leq \|A\|_M + \|B\|_M, \quad \forall A, B \in \mathbb{M}^n(\mathbb{C});$

(4) $\|Ax\|_V \leq \|A\|_M \|x\|_V, \quad \forall A \in \mathbb{M}^n(\mathbb{C}), \;\; \forall x \in \mathbb{C}^n;$

(5) $\|AB\|_M \leq \|A\|_M \|B\|_M, \quad \forall A, B \in \mathbb{M}^n(\mathbb{C}).$

Dem.: $(\cdots)$

Proposição. Sendo M a norma associada à norma vectorial V e I a matriz identidade, então $\|I\|_M = 1$.

Exemplo. Normas matriciais em $\mathbb{M}^n(\mathbb{C})$ associadas a normas vectoriais em $\mathbb{C}^n$:

$x = [x_i] \in \mathbb{C}^n$	$A = [a_{ij}] \in \mathbb{M}^n(\mathbb{C})$
$\|x\|_1 = \sum_{i=1}^{n} \|x_i\|$	$\|A\|_1 = \max_{1\leq j\leq n} \sum_{i=1}^{n} \|a_{ij}\|$ **norma por colunas**
$\|x\|_2 = \sqrt{\sum_{i=1}^{n} \|x_i\|^2}$	$\|A\|_2 = \sqrt{r_\sigma(A^*A)}$ **norma Euclideana**
$\|x\|_\infty = \max_{1\leq i\leq n} \|x_i\|$	$\|A\|_\infty = \max_{1\leq i\leq n} \sum_{j=1}^{n} \|a_{ij}\|$ **norma por linhas**

Nota. A^* designa a matriz transposta conjugada da matriz A.

Definição. Seja $A \in \mathbb{M}^n(\mathbb{C})$ e seja $\sigma(A)$ o **espectro** de A, isto é, o conjunto dos valores próprios de A. Chama-se **raio espectral** de A e representa-se por $r_\sigma(A)$ a

$$r_\sigma(A) = \max_{\lambda\in\sigma(A)} |\lambda|.$$

Proposição. A norma matricial associada à norma da soma em $\mathbb{C}^n$ tem a forma

$$\|A\|_1 = \max_{1\leq j\leq n} \sum_{i=1}^{n} |a_{ij}|.$$

Dem.: $(\cdots)$

Proposição. A norma matricial associada à norma do máximo em $\mathbb{C}^n$ tem a forma

$$\|A\|_\infty = \max_{1\leq i\leq n} \sum_{j=1}^{n} |a_{ij}|.$$

Dem.: (Exercício)

Proposição. A norma matricial associada à norma Euclidiana em $\mathbb{C}^n$ tem a forma

$$\|A\|_2 = \sqrt{r_\sigma(A^*A)}.$$

Nota. O cálculo de $\|A\|_2$ é muito mais complicado do que o de $\|A\|_1$ ou $\|A\|_\infty$. Se apenas precisarmos de uma estimativa de $\|A\|_2$ podemos recorrer às desigualdades seguintes.

Proposição. Sendo $A \in \mathbb{M}^n(\mathbb{C})$, então:

(1) $\frac{1}{\sqrt{n}} \|A\|_1 \leq \|A\|_2 \leq \sqrt{n}\, \|A\|_1$

(2) $\frac{1}{\sqrt{n}} \|A\|_\infty \leq \|A\|_2 \leq \sqrt{n}\, \|A\|_\infty$

(3) $\|A\|_2 \leq \sqrt{\|A\|_1 \, \|A\|_\infty}$

(4) $\frac{1}{\sqrt{n}} \|A\|_F \leq \|A\|_2 \leq \|A\|_F$.

Dem.: (Exercício)

Nota. Sendo $A \in \mathbb{M}^n(\mathbb{C})$, $A = [a_{ij}]$, define-se a **norma de Frobenius** de A por

$$\|A\|_F = \left[\sum_{i,j=1}^{n} |a_{ij}|^2 \right]^{1/2}.$$

Esta norma é regular, isto é,

$$\|AB\|_F \leq \|A\|_F \|B\|_F, \qquad \forall A, B \in \mathbb{M}^n(\mathbb{C}).$$

Esta norma é compatível com a norma euclideana para vectores em $\mathbb{C}^n$, isto é,

$$\|Ax\|_2 \leq \|A\|_F \|x\|_2, \qquad \forall A \in \mathbb{M}^n(\mathbb{C}), \ \ \forall x \in \mathbb{C}^n.$$

Esta norma <u>não</u> está associada a nenhuma norma vectorial pois $\|I\|_F = \sqrt{n}$.

Exemplo. Sendo

$$A = \begin{bmatrix} 2 & 1 & 1 \\ -1 & 3 & 1 \\ 1 & -2 & 2 \end{bmatrix}$$

calcular $\|A\|_1, \|A\|_\infty, \|A\|_F, r_\sigma(A), \|A\|_2$.

Resolução:

$\|A\|_1 = \max\{4, 6, 4\} = 6$

$\|A\|_\infty = \max\{4, 5, 5\} = 5$

$\|A\|_F = (4 + 1 + 1 + 1 + 9 + 1 + 1 + 4 + 4)^{1/2} = \sqrt{26} = 5.09902$

$r_\sigma(A) = \max\limits_{\lambda \in \sigma(A)} |\lambda|$

$$\begin{aligned}\sigma(A) &= \{\lambda \in \mathbb{C} : \ -\lambda^3 + 7\lambda^2 - 18\lambda + 18 = 0\} \\ &= \{3, \ 2 + i\sqrt{2}, \ 2 - i\sqrt{2}\}\end{aligned}$$

$$r_\sigma(A) = 3$$

$$\|A\|_2 = \sqrt{r_\sigma(A^T A)}$$

$$\begin{aligned}\sigma(A^T A) &= \{\lambda \in \mathbb{C} : \ -\lambda^3 + 26\lambda^2 - 186\lambda + 324 = 0\} \\ &= \{2.58018, \ 8.31148, \ 15.1083\}\end{aligned}$$

$$\|A\|_2 = 3.88695$$

Proposição. Seja $A \in \mathbb{M}^n(\mathbb{C})$. Então:

(1) Qualquer que seja a norma matricial M associada a uma certa norma vectorial em $\mathbb{C}^n$, verifica-se
$$r_\sigma(A) \leq \|A\|_M.$$

(2) Qualquer que seja $\varepsilon > 0$ existe uma norma $N(\varepsilon)$ associada a uma certa norma vectorial em $\mathbb{C}^n$ tal que
$$\|A\|_{N(\varepsilon)} \leq r_\sigma(A) + \varepsilon.$$

Dem.: (1) $(\cdots)$

Corolário. Seja $A \in \mathbb{M}^n(\mathbb{C})$. Então $r_\sigma(A) < 1$ se e só se $\|A\|_M < 1$ para alguma norma matricial M associada a uma norma vectorial em $\mathbb{C}^n$.

Dem.: $(\cdots)$

Nota. O raio espectral de A é pois o ínfimo de todas as normas matriciais de A associadas a normas vectoriais em $\mathbb{C}^n$.

Condicionamento de sistemas lineares

• Consideremos o problema de determinar a solução de um sistema linear

$$Ax = b,$$

onde $A \in \mathbb{M}^n(\mathbb{C}), \det A \neq 0, b \in \mathbb{C}^n$. Os dados do problema são neste caso a matriz A e o vector b. A solução é o vector $x = A^{-1}b$. Pretendemos analisar a forma como os erros nos dados afectam os erros na solução. Para isso consideramos um outro sistema linear

$$\tilde{A}\tilde{x} = \tilde{b},$$

e vamos procurar exprimir o erro relativo da solução do segundo sistema em relação à do primeiro, $\dfrac{x - \tilde{x}}{\|x\|}$, em termos dos erros relativos dos dados do segundo sistema em relação aos do primeiro, isto é, $\dfrac{A - \tilde{A}}{\|A\|}$ e $\dfrac{b - \tilde{b}}{\|b\|}$.

Proposição. Consideremos os sistemas lineares

$$Ax = b, \qquad A\tilde{x} = \tilde{b},$$

onde $b, \tilde{b} \in \mathbb{C}^n$, $A \in \mathbb{M}^n(\mathbb{C})$, $\det A \neq 0$. Então

$$\frac{\|\delta_{\tilde{b}}\|_V}{\mathrm{cond}_M(A)} \leq \|\delta_{\tilde{x}}\|_V \leq \mathrm{cond}_M(A)\, \|\delta_{\tilde{b}}\|_V$$

onde

$$\delta_{\tilde{x}} = \frac{x - \tilde{x}}{\|x\|_V}, \qquad \delta_{\tilde{b}} = \frac{b - \tilde{b}}{\|b\|_V}, \qquad \mathrm{cond}_M(A) = \|A\|_M \|A^{-1}\|_M.$$

M designa a norma matricial associada à norma vectorial V.

Dem.: $(\cdots)$

Proposição. Consideremos os sistemas lineares

$$Ax = b, \qquad \tilde{A}\tilde{x} = \tilde{b},$$

onde $b, \tilde{b} \in \mathbb{C}^n$, $A, \tilde{A} \in \mathbb{M}^n(\mathbb{C})$, $\det A \neq 0$ e $\tilde{A}$ é tal que

$$\|A - \tilde{A}\|_M \|A^{-1}\|_M < 1.$$

Então $\tilde{A}$ é não singular e verifica-se a desigualdade

$$\|\delta_{\tilde{x}}\|_V \leq \frac{\mathrm{cond}_M(A)}{1 - \|\delta_{\tilde{A}}\|_M \,\mathrm{cond}_M(A)} \left(\|\delta_{\tilde{A}}\|_M + \|\delta_{\tilde{b}}\|_V\right),$$

onde

$$\delta_{\tilde{x}} = \frac{x - \tilde{x}}{\|x\|_V}, \qquad \delta_{\tilde{b}} = \frac{b - \tilde{b}}{\|b\|_V}, \qquad \delta_{\tilde{A}} = \frac{A - \tilde{A}}{\|A\|_M}, \qquad \mathrm{cond}_M(A) = \|A\|_M \|A^{-1}\|_M.$$

M designa a norma matricial associada à norma vectorial V.

Definição. Seja $A \in \mathbb{M}^n(\mathbb{C})$ uma matriz não singular. Chama-se **número de condição de A relativamente à norma M** a

$$\mathrm{cond}_M(A) = \|A\|_M \|A^{-1}\|_M.$$

Chama-se **número de condição de A relativamente ao raio espectral** a

$$\mathrm{cond}_*(A) = r_\sigma(A) r_\sigma(A^{-1}).$$

Nota. O número de condição de A relativamente à norma $p \geq 1$ é designado por $\mathrm{cond}_p(A)$.

Proposição. Seja $A \in \mathbb{M}^n(\mathbb{C})$ uma matriz não singular. Então:

(1) $\text{cond}_*(A) = \dfrac{\max\limits_{\lambda\in\sigma(A)} |\lambda|}{\min\limits_{\lambda\in\sigma(A)} |\lambda|}$;

(2) $\text{cond}_p(A) \geq \text{cond}_*(A) \geq 1, \quad \forall p \geq 1$;

(3) $A = A^* \;\Rightarrow\; \text{cond}_2(A) = \text{cond}_*(A)$;

(4) $AA^* = I \;\Rightarrow\; \text{cond}_2(A) = \text{cond}_*(A) = 1.$

Dem.: $(\cdots)$

Nota. (1) Uma matriz $A \in \mathbb{M}^n(\mathbb{C})$ tal que $A^* = A$ diz-se uma matriz **hermitiana**. Uma matriz hermitiana real é uma matriz **simétrica.**

(2) Uma matriz $A \in \mathbb{M}^n(\mathbb{C})$ tal que $A^*A = AA^* = I$ diz-se uma matriz **unitária**. Uma matriz unitária real é uma matriz **ortogonal.**

Definição. Um sistema linear com matriz A diz-se **bem condicionado** se $\text{cond}_p(A) \overset{\approx}{>} 1$ e **mal condicionado** se $\text{cond}_p(A) \gg 1$.

Exemplo. Sendo

$$A = \begin{bmatrix} 2 & 1 & 1 \\ -1 & 3 & 1 \\ 1 & -2 & 2 \end{bmatrix}$$

calcular $\text{cond}_1(A), \text{cond}_2(A), \text{cond}_\infty(A), \text{cond}_*(A)$.

Resolução:

$$A^{-1} = \frac{1}{18}\begin{bmatrix} 8 & -4 & -2 \\ 3 & 3 & -3 \\ -1 & 5 & 7 \end{bmatrix}$$

$$\|A^{-1}\|_1 = \frac{1}{18}\max\{12, 12, 12\} = \frac{2}{3}$$

$$\|A^{-1}\|_\infty = \frac{1}{18}\max\{14, 9, 13\} = \frac{7}{9}$$

$$r_\sigma(A^{-1}) = \frac{1}{\min_{\lambda\in\sigma(A)}|\lambda|} = \frac{\sqrt{6}}{6}$$

$$r_\sigma\left((A^{-1})^T A^{-1}\right) = \frac{1}{\min_{\lambda\in\sigma(A^TA)}|\lambda|} = 0.387570$$

$$\|A^{-1}\|_2 = 0.622551$$

$$\text{cond}_1(A) = 6 \times \frac{2}{3} = 4$$

$$\text{cond}_2(A) = 3.88695 \times 0.622551 = 2.41982$$

$$\text{cond}_\infty(A) = 5 \times \frac{7}{9} = 3.88889$$

$$\text{cond}_*(A) = 3 \times \frac{\sqrt{6}}{6} = 1.22474$$

Exemplo. Matriz de Hilbert,

$$H_n \in \mathbb{M}^n(\mathbb{R}), \qquad (H_n)_{ij} = \frac{1}{i+j-1}.$$

◇ A matriz inversa é conhecida explicitamente:

$$\left(H_n^{-1}\right)_{ij} = \frac{(-1)^{i+j}(n+i-1)!(n+j-1)!}{(i+j-1)[(i-1)!(j-1)!]^2(n-i)!(n-j)!}$$

◇ A matriz aparece, por exemplo, na aproximação mínimos quadrados de uma função.

◇ A matriz é muito mal condicionada:

n	$\text{cond}_*(H_n)$
2	1.93×10
3	5.24×10^2
4	1.55×10^4
5	4.77×10^5
6	1.50×10^7
7	4.75×10^8
8	1.53×10^{10}
9	4.93×10^{11}
10	1.60×10^{13}

◇ A matriz é utilizada para testar programas para resolver sistemas lineares mal condicionados.

Exemplo. Considerem-se os sistemas

$$H_n x = b_n, \qquad n = 2, 3, 4, 5 \tag{1}$$

onde H_n é a matriz de Hilbert de ordem n e $b_n \in \mathbb{C}^n$ é um vector cujas componentes são todas iguais a 1. Considerem-se também os sistemas

$$\tilde{H}_n \tilde{x} = b_n, \qquad n = 2, 3, 4, 5, \tag{2}$$

onde $\tilde{H}_n$ é a matriz que representa a matrix de Hilbert H_n num sistema de ponto flutuante com 4 dígitos na mantissa. Determinar os erros relativos das soluções dos sistemas (2) em relação aos sistemas (1), com o mesmo valor de n, e interpretar os resultados à luz da desigualdade do teorema anterior.

n	$\|\delta_{\tilde{H}_n}\|_\infty$	$\|\delta_{\tilde{x}}\|_\infty$	$\text{cond}_\infty(H_n)$	$z_n = \|\delta_{\tilde{H}_n}\|_\infty \times \text{cond}_\infty(H_n)$	$\frac{z_n}{1-z_n}$
2	0.222×10^{-4}	0.400×10^{-3}	27	0.600×10^{-3}	0.600×10^{-3}
3	0.182×10^{-4}	0.698×10^{-2}	748	0.136×10^{-1}	0.138×10^{-1}
4	0.366×10^{-4}	0.269	28375	0.104×10^{1}	
5	0.480×10^{-4}	0.914	943656	0.453×10^{2}	

Resolução numérica de sistemas lineares

• *Métodos directos*: a solução exacta (na ausência de erros de arredondamento) é obtida num número finito de passos. São exemplos: o *método de eliminação de Gauss*, de que trataremos brevemente para introduzir a *pesquisa parcial de pivot*; os *métodos de factorização*, a que apenas faremos uma breve referência.

• *Métodos iterativos*: a solução aproximada converge para a solução exacta quando o número de iteradas tende para infinito. São exemplos, de que trataremos seguidamente: *método de Jacobi, método de Gauss-Seidel, método de Jacobi modificado ou com relaxação, método de Gauss-Seidel modificado ou com relaxação ou método SOR.*

Método de eliminação de Gauss com pesquisa parcial de pivot

Definição (pesquisa parcial de pivot). No passo k da eliminação de Gauss considere-se

$$s = \min \left\{ r : \left| a_{rk}^{(k)} \right| = \max_{k \le i \le n} \left| a_{ik}^{(k)} \right| \right\}.$$

s é pois o menor dos índices das linhas $r \geq k$ para os quais é atingido o valor máximo dos valores absolutos dos elementos $\left| a_{ik}^{(k)} \right|$ da coluna k localizados abaixo e sobre a diagonal principal. Se $s > k$ trocam-se as linhas s e k da matriz $A^{(k)}$ e do vector $b^{(k)}$; e prossegue-se com o passo k da eliminação de Gauss. Esta estratégia implica que

$$|m_{ik}| \leq 1, \qquad i = k+1, \ldots, n.$$

(0) $A^{(1)} = A, \quad b^{(1)} = b$

$$\left[\begin{array}{l} i = 1, 2, \ldots, n \\ \quad \left[\begin{array}{l} j = 1, 2, \ldots, n \\ \quad a_{ij}^{(1)} = a_{ij} \end{array}\right. \\ \quad b_i^{(1)} = b_i \end{array}\right.$$

(1) Redução da matriz A à forma triangular superior

$A^{(1)} \to A^{(2)} \to \cdots \to A^{(n)}$,

(2) Transformação do 2^{o} membro do sitema

$b^{(1)} \to b^{(2)} \to \cdots \to b^{(n)}$

$$\left[\begin{array}{l} k = 1, 2, \ldots, n-1 \qquad \left(A^{(k)} \to A^{(k+1)}, \quad b^{(k)} \to b^{(k+1)}\right) \\ \quad \left[\begin{array}{l} s > k, \quad s = \min\left\{ r : \left|a_{rk}^{(k)}\right| = \max\limits_{k \le i \le n} \left|a_{ik}^{(k)}\right| \right\}. \\ \quad \left[\begin{array}{l} j = k, k+1, \ldots, n \\ \quad a_{kj}^{(k)} \leftrightarrow a_{sj}^{(k)} \end{array}\right. \\ \quad b_k^{(k)} \leftrightarrow b_s^{(k)} \end{array}\right. \\ \quad \left[\begin{array}{l} i = k+1, k+2, \ldots, n \\ \quad m_{ik} = \dfrac{a_{ik}^{(k)}}{a_{kk}^{(k)}} \\ \quad \left[\begin{array}{l} j = k+1, k+2, \ldots, n \\ \quad a_{ij}^{(k+1)} = a_{ij}^{(k)} - m_{ik} a_{kj}^{(k)} \end{array}\right. \\ \quad b_i^{(k+1)} = b_i^{(k)} - m_{ik} b_k^{(k)} \end{array}\right. \end{array}\right.$$

(3) Resolução do sistema transformado : $\quad A^{(n)} x = b^{(n)}$

$$\left[\begin{array}{l} x_n = \dfrac{b_n^{(n)}}{a_{nn}^{(n)}} \\ \quad \left[\begin{array}{l} i = n-1, n-2, \ldots, 1 \\ \quad x_i = \dfrac{1}{a_{ii}^{(i)}} \left(b_i^{(i)} - \displaystyle\sum_{j=i+1}^{n} a_{ij}^{(i)} x_j \right) \end{array}\right. \end{array}\right.$$

Nota. Utilização do método de eliminação de Gauss para calcular o determinante de uma matriz $A \in \mathbb{M}^n(\mathbb{C})$, não singular:

$$\det A = (-1)^\nu \det A^{(n)} = (-1)^\nu a_{11}^{(n)} a_{22}^{(n)} \cdots a_{nn}^{(n)},$$

onde ν designa o número de troca de linhas efectuado. Comparemos, do ponto de vista da eficiência computacional, medida pelo número de multiplicações e divisões necessárias para obter o resultado, este método com o método baseado na utilização directa da definição:

$$\det A = \sum_{p \in \Pi} \operatorname{sgn}(p) a_{1\alpha} a_{2\beta} \ldots a_{n\omega},$$

onde Π designa o conjunto de todas as permutações de $(1, 2, \ldots, n)$, $p = (\alpha, \beta, \ldots, \omega)$ e $\operatorname{sgn}(p) = \pm 1$.

n	$N_{\text{def}}(n) = n!(n-1)$	$N_{EG}(n) = \frac{1}{3}(n-1)(n^2+n+3)$	$\tilde{N}_{EG}(n) = \frac{n^3}{3}$
2	2	3	
10	3.27×10^7	339	333
100	9.24×10^{159}	333399	333333

$N_{\text{def}}(n)$ designa o número de multiplicações necessárias para calcular $\det A$ por definição, $N_{EG}(n)$ designa o número de multiplicações e divisões necessárias para obter $\det A$ usando o método de eliminação de Gauss para obter a matriz $A^{(n)}$ e $\tilde{N}_{EG}(n)$ designa o valor aproximado (assimptótico) de $N_{EG}(n)$ para valores elevados de n.

Nota. Utilização do método de eliminação de Gauss para calcular a matriz inversa A^{-1} de uma matriz $A \in \mathbb{M}^n(\mathbb{C})$, não singular. Sendo

$$AA^{-1} = I,$$

e designando por $c_1, \ldots, c_n$ as colunas de A^{-1} e por $e_1, \ldots, e_n$ as colunas da matriz identidade, podemos escrever

$$A[c_1 \;\; c_2 \;\; \cdots \;\; c_n] = [e_1 \;\; e_2 \;\; \cdots \;\; e_n].$$

A matriz A^{-1} pode pois ser obtida determinando as soluções $c_1, \ldots, c_n$ dos n sistemas $Ac_1 = e_1, \ldots, Ac_n = e_n$ por eliminação de Gauss. Note-se que todos estes sistemas têm a mesma matriz A pelo que a parte de redução da matriz à forma triangular superior é comum a todos eles.

Nota. Utilização do método de eliminação de Gauss para obter a **factorização triangular** de uma matriz $A \in \mathbb{M}^n(\mathbb{C})$. Sendo A uma matriz não singular então existe uma matriz de permutação P tal que a matriz PA admite uma única factorização triangular $PA = LU$, onde L é uma matriz triangular inferior com diagonal principal unitária e U é uma matriz triangular superior com todos os elementos na diagonal principal diferentes de zero. Os elementos de L por baixo da diagonal principal na coluna k são os

multiplicadores utilizados no passo k da eliminação de Gauss, isto é,

$$L = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ m_{21} & 1 & 0 & \cdots & 0 & 0 \\ m_{31} & m_{32} & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ m_{n1} & m_{n2} & m_{n3} & \cdots & m_{n,n-1} & 1 \end{bmatrix}$$

A matriz U é a matriz do sistema após a eliminação de Gauss, isto é,

$$U = A^{(n)}.$$

A factorização triangular de uma matriz permite reduzir a resolução de um qualquer sistema linear $Ax = b$ à resolução de dois sistemas lineares com matrizes triangulares, a qual se faz facilmente:

$$Ax = b \;\Leftrightarrow\; LUx = Pb \;\Leftrightarrow\; \begin{cases} Ly = Pb \\ Ux = y \end{cases}$$

Exemplo. Considere-se o sistema linear

$$\begin{bmatrix} 1 & 5 \\ 500 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 5 \\ 2 \end{bmatrix}.$$

(a) Determinar a solução exacta do sistema pelo método de eliminação de Gauss.

(b) Supondo que se efectuam os cálculos num sistema FP(10, 3, -10, 10), com arredondamento simétrico, determinar a solução aproximada do sistema pelo método de eliminação de Gauss.

(c) Idem, pelo método de eliminação de Gauss com pesquisa parcial de pivot.

(d) Comparar os erros relativos das soluções obtidas nas alíneas anteriores.

Resolução:

(a)

$$\begin{bmatrix} 1 & 5 \\ 500 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 5 \\ 2 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 5 \\ 0 & -2499 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 5 \\ -2498 \end{bmatrix}$$

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \dfrac{5}{2499} \\ \dfrac{2498}{2499} \end{bmatrix} = \begin{bmatrix} 0.200080032\cdots \times 10^{-2} \\ 0.999599840\cdots \end{bmatrix}$$

(b)

$$\begin{bmatrix} 1.00 & 5.00 \\ 500. & 1.00 \end{bmatrix} \begin{bmatrix} \tilde{x} \\ \tilde{y} \end{bmatrix} = \begin{bmatrix} 5.00 \\ 2.00 \end{bmatrix}$$

$$\begin{bmatrix} 1.00 & 5.00 \\ 0.00 & -2500. \end{bmatrix} \begin{bmatrix} \tilde{x} \\ \tilde{y} \end{bmatrix} = \begin{bmatrix} 5.00 \\ -2500. \end{bmatrix}$$

$$\begin{bmatrix} \tilde{x} \\ \tilde{y} \end{bmatrix} = \begin{bmatrix} 0.00 \\ 1.00 \end{bmatrix}$$

(c)

$$\begin{bmatrix} 500. & 1.00 \\ 1.00 & 5.00 \end{bmatrix} \begin{bmatrix} \tilde{x} \\ \tilde{y} \end{bmatrix} = \begin{bmatrix} 2.00 \\ 5.00 \end{bmatrix}$$

$$\begin{bmatrix} 500. & 1.00 \\ 0.00 & 5.00 \end{bmatrix} \begin{bmatrix} \tilde{x} \\ \tilde{y} \end{bmatrix} = \begin{bmatrix} 2.00 \\ 5.00 \end{bmatrix}$$

$$\begin{bmatrix} \tilde{x} \\ \tilde{y} \end{bmatrix} = \begin{bmatrix} 0.00200 \\ 1.00 \end{bmatrix}$$

(d) (b) $|\delta_{\tilde{x}}| = 1.$, $|\delta_{\tilde{y}}| = 0.0004$

(c) $|\delta_{\tilde{x}}| = 0.0004$, $|\delta_{\tilde{y}}| = 0.0004$

Métodos iterativos

• Vamos considerar métodos iterativos para obter valores aproximados da solução do sistema linear $Ax = b$, $A \in \mathbb{M}^n(\mathbb{C}), b \in \mathbb{C}^n$, da forma

$$\begin{cases} x^{(k+1)} = Cx^{(k)} + w, & k \in \mathbb{N}, \\ x^{(0)} = \xi_0 \in \mathbb{C}^n. \end{cases}$$

Tratam-se pois de métodos iterativos a um passo com função iteradora $\Phi : \mathbb{C}^n \to \mathbb{C}^n$ definida por $\Phi(x) = Cx + w$, onde $C \in \mathbb{M}^n(\mathbb{C})$ e $w \in \mathbb{C}^n$. C é a **matriz iteradora** do método.

Definição. O método iterativo com função iteradora Φ diz-se **consistente** com o sistema $Ax = b$ se este sistema e o sistema $x = \Phi(x)$ tiverem a mesma solução.

Proposição. O método iterativo com função iteradora Φ é consistente com o sistema $Ax = b$ se e só se

$$(I - C)A^{-1}b = w.$$

Proposição. O método iterativo com função iteradora Φ tal que

$$C = -M^{-1}N = I - M^{-1}A, \qquad w = M^{-1}b,$$

onde $M, N \in \mathbb{M}^n(\mathbb{C})$, M invertível, $M + N = A$, é consistente.

Dem.:

$$Ax = b \;\Rightarrow\; (M+N)x = b \;\Rightarrow\; Mx = -Nx + b \;\Rightarrow\; x = -M^{-1}Nx + M^{-1}b$$

Nota. A forma "intermédia" do método iterativo

$$Mx^{(k+1)} = -Nx^{(k)} + b, \qquad k \in \mathbb{N},$$

será útil mais adiante.

Proposição. Qualquer matriz $A \in \mathbb{M}^n(\mathbb{C})$, $A = [a_{ij}]$, pode ser decomposta na soma

$$A = L_A + D_A + U_A,$$

onde

(i) $L_A = [l_{ij}] \in \mathbb{M}^n(\mathbb{C})$ é uma matriz estritamente triangular inferior com elementos

$$l_{ij} = a_{ij}, \quad i > j, \qquad l_{ij} = 0, \quad i \leq j.$$

(ii) $D_A = [d_{ij}] \in \mathbb{M}^n(\mathbb{C})$ é uma matriz diagonal com elementos

$$d_{ij} = a_{ij}, \quad i = j, \qquad d_{ij} = 0, \quad i \neq j.$$

(iii) $U_A = [u_{ij}] \in \mathbb{M}^n(\mathbb{C})$ é uma matriz estritamente triangular superior com elementos

$$u_{ij} = a_{ij}, \quad i < j, \qquad u_{ij} = 0, \quad i \geq j.$$

Definição. O **método de Jacobi** corresponde a tomar

$$M = D, \qquad N = L + U,$$

na decomposição de $A = L + D + U = M + N$, exigindo que D seja invertível. Toma pois a forma:

$$\begin{cases} x^{(k+1)} = D^{-1}[b - (L+U)x^{(k)}], & k \in \mathbb{N}, \\ x^{(0)} = \xi_0 \in \mathbb{C}^n, \end{cases}$$

ou, componente a componente,

$$\begin{cases} x_i^{(k+1)} = \dfrac{1}{a_{ii}} \left[b_i - \displaystyle\sum_{\substack{j=1 \\ j \neq i}}^{n} a_{ij} x_j^{(k)} \right], & i = 1, \ldots, n, \qquad k \in \mathbb{N}, \\ x_i^{(0)} = \xi_{0i} \in \mathbb{C}, & i = 1, \ldots, n. \end{cases}$$

A matriz iteradora é

$$C_J = -D^{-1}(L+U) = I - D^{-1}A.$$

Nota. O método de Jacobi é também conhecido por *método das substituições simultâneas.*

Definição. O **método de Gauss-Seidel** corresponde a tomar

$$M = D + L, \qquad N = U,$$

na decomposição de $A = L + D + U = M + N$, exigindo que $D + L$ seja invertível. Toma pois a forma;

$$\begin{cases} x^{(k+1)} = D^{-1}\left[b - Lx^{(k+1)} - Ux^{(k)}\right], & k \in \mathbb{N}, \\ x^{(0)} = \xi_0 \in \mathbb{C}^n, \end{cases}$$

ou, componente a componente,

$$\begin{cases} x_i^{(k+1)} = \dfrac{1}{a_{ii}}\left[b_i - \displaystyle\sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=1+1}^{n} a_{ij}x_j^{(k)}\right], & i = 1, \ldots, n, \quad k \in \mathbb{N}, \\ x_i^{(0)} = \xi_{0i} \in \mathbb{C}, \quad i = 1, \ldots, n. \end{cases}$$

A matriz iteradora é

$$C_{GS} = -(D + L)^{-1}U = I - (D + L)^{-1}A.$$

Nota. O método de Gauss-Seidel é também conhecido por *método das substituições sucessivas.*

Definição. O **método de Jacobi modificado** ou **método de Jacobi com relaxação** corresponde a tomar

$$M = \frac{D}{\omega}, \qquad N = \left(1 - \frac{1}{\omega}\right) D + L + U,$$

na decomposição de $A = L + D + U = M + N$, exigindo que D seja invertível. ω é um parâmetro real positivo conhecido por **parâmetro de relaxação**. O método toma pois a forma:

$$\begin{cases} x^{(k+1)} = (1 - \omega)x^{(k)} + \omega D^{-1}[b - (L + U)x^{(k)}], & k \in \mathbb{N}, \\ x^{(0)} = \xi_0 \in \mathbb{C}^n, \end{cases}$$

ou, componente a componente,

$$\begin{cases} x_i^{(k+1)} = (1 - \omega)x_i^{(k)} + \dfrac{\omega}{a_{ii}}\left[b_i - \displaystyle\sum_{\substack{j=1 \\ j \neq i}}^{n} a_{ij}x_j^{(k)}\right], & i = 1, \ldots, n, \quad k \in \mathbb{N}, \\ x_i^{(0)} = \xi_{0i} \in \mathbb{C}, \quad i = 1, \ldots, n. \end{cases}$$

A matriz iteradora é

$$C_{Jm}(\omega) = I - \omega D^{-1}A.$$

Nota. Para $\omega = 1$ o método reduz-se ao método de Jacobi.

Definição. O **método de Gauss-Seidel modificado** ou ou **método de Gauss-Seidel com relaxação** ou **método SOR** corresponde a tomar

$$M = \frac{D}{\omega} + L, \qquad N = \left(1 - \frac{1}{\omega}\right) D + U,$$

na decomposição de $A = L + D + U = M + N$, exigindo que $D/\omega + L$ seja invertível. ω é um parâmetro real positivo conhecido por **parâmetro de relaxação**. O método SOR toma pois a forma;

$$\begin{cases} x^{(k+1)} = (1-\omega)x^{(k)} + \omega D^{-1}\left[b - Lx^{(k+1)} - Ux^{(k)}\right], & k \in \mathbb{N}, \\ x^{(0)} = \xi_0 \in \mathbb{C}^n, \end{cases}$$

ou, componente a componente,

$$\begin{cases} x_i^{(k+1)} = (1-\omega)x_i^{(k)} + \dfrac{\omega}{a_{ii}}\left[b_i - \displaystyle\sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=1+1}^{n} a_{ij}x_j^{(k)}\right], \quad i = 1,\ldots,n, \quad k \in \mathbb{N}, \\ x_i^{(0)} = \xi_{0i} \in \mathbb{C}, \qquad i = 1,\ldots,n. \end{cases}$$

A matriz iteradora é

$$C_{SOR}(\omega) = I - \omega(D + \omega L)^{-1}A.$$

Nota. SOR é o acrónimo de *successive over-relaxation*. Para $\omega = 1$ o método SOR reduz-se ao método de Gauss-Seidel.

Nota. A introdução do parâmetro de relaxação poderá permitir tornar convergente um método divergente ou acelerar a convergência de um método convergente.

Convergência dos métodos iterativos

• Consideremos o método iterativo consistente com o sistema linear $Ax = b$, $A = M + N$,

$$\begin{cases} x^{(k+1)} = Cx^{(k)} + w, & k \in \mathbb{N}, \\ x^{(0)} = \xi_0 \in \mathbb{C}^n, \end{cases} \qquad (\#)$$

onde

$$C = -M^{-1}N = I - M^{-1}A, \qquad w = M^{-1}b.$$

Proposição. O método iterativo (#) converge para a solução x do sistema $Ax = b$, $\forall x^{(0)} \in \mathbb{C}^n$, isto é,

$$\lim_{k\to\infty} x^{(k)} = x, \qquad \forall x^{(0)} \in \mathbb{C}^n,$$

ou

$$\lim_{k\to\infty} e^{(k)} = 0, \qquad \forall x^{(0)} \in \mathbb{C}^n,$$

onde $e^{(k)} = x - x^{(k)}$, se e só se

$$\lim_{k\to\infty} C^k e^{(0)} = 0, \qquad \forall e^{(0)} \in \mathbb{C}^n.$$

Dem.: De $x = \Phi(x)$ e $x^{(k+1)} = \Phi(x^{(k)})$ obtém-se por subtracção $e^{(k+1)} = Ce^{(k)}$ e por indução $e^{(k)} = C^k e^{(0)}$.

Proposição. O método iterativo (#) converge para a solução x do sistema $Ax = b$, $\forall x^{(0)} \in \mathbb{C}^n$, se existir uma norma matricial M, associada a uma certa norma vectorial V em $\mathbb{C}^n$, tal que $\|C\|_M < 1$. Neste caso são válidas as seguintes estimativas de erro:

(1) $\|e^{(k+1)}\|_V \leq c\|e^{(k)}\|_V$;

(2) $\|e^{(k)}\|_V \leq c^k\|e^{(0)}\|_V$;

(3) $\|e^{(k)}\|_V \leq \dfrac{1}{1-c}\,\|x^{(k+1)} - x^{(k)}\|_V$;

(4) $\|e^{(k+1)}\|_V \leq \dfrac{c}{1-c}\,\|x^{(k+1)} - x^{(k)}\|_V$;

(5) $\|e^{(k)}\|_V \leq \dfrac{c^k}{1-c}\,\|x^{(1)} - x^{(0)}\|_V$;

onde $e^{(k)} = x - x^{(k)}$, $k \in \mathbb{N}$, e $c = \|C\|_M$.

Dem.: $(\cdots)$

Nota. A estimativa "a posteriori"(4) pode ser usada como critério de paragem do método iterativo:

$$\|e^{(k+1)}\|_V \leq \frac{c}{1-c}\,\|x^{(k+1)} - x^{(k)}\|_V < \varepsilon$$

$$\|x^{(k+1)} - x^{(k)}\|_V < \varepsilon\left(\frac{1}{c} - 1\right).$$

Repare-se que

$$\begin{cases} 1 \leq \dfrac{1}{c} - 1, & 0 < c \leq \dfrac{1}{2}, \\ 0 < \dfrac{1}{c} - 1 \leq 1, & \dfrac{1}{2} \leq c < 1. \end{cases}$$

Nota. Todas estas estimativas permanecem válidas se substituirmos c por uma outra quantidade $\tilde{c}$ tal que $c < \tilde{c} < 1$. Em particular, $\|C\|_2$, que é difícil de calcular, pode ser substituída por $\|C\|_F < 1$. Recorde-se que $\|C\|_2 \leq \|C\|_F$, $\forall C \in \mathbb{M}^n(\mathbb{C})$.

Proposição. O método iterativo (#) converge para a solução x do sistema $Ax = b$, $\forall x^{(0)} \in \mathbb{C}^n$, se e só se $r_\sigma(C) < 1$.

Dem.: $(\cdots)$

Nota. Estes dois teoremas, conjugados com um teorema anterior que nos diz que o raio espectral de uma matriz é o ínfimo de todas as normas associadas a normas vectoriais da matriz, sugerem que o raio espectral determina a rapidez de convergência do método, sendo a convergência tanto mais rápida quanto menor for o raio espectral da matriz iteradora. Mais precisamente pode mostrar-se que:

Proposição. Para o método iterativo (#) os erros $e^{(k)} = x - x^{(k)}$ satisfazem a

$$\sup_{e^{(0)} \in \mathbb{C}^n \setminus \{0\}} \limsup_{k \to \infty} \left(\frac{\|e^{(k)}\|_V}{\|e^{(0)}\|_V} \right)^{1/k} = r_\sigma(C),$$

onde V designa qualquer norma em $\mathbb{C}^n$.

Nota.

(a) Este resultado significa que para valores de k elevados, podemos escrever em muitos casos

$$\|e^{(k+1)}\| \approx r_\sigma(C) \|e^{(k)}\|.$$

(b) Partindo de $x^{(k+1)} = Cx^{(k)} + w$ e de $x^{(k)} = Cx^{(k-1)} + w$ obtém-se

$$x^{(k+1)} - x^{(k)} = C\left(x^{(k)} - x^{(k-1)}\right),$$

isto é, a diferença de duas iteradas consecutivas satisfaz à mesma igualdade que os erros de duas iteradas consecutivas:

$$e^{(k+1)} = Ce^{(k)}.$$

Para valores de k elevados verifica-se também

$$\|x^{(k+1)} - x^{(k)}\| \approx r_\sigma(C) \|x^{(k)} - x^{(k-1)}\|.$$

(c) Na prática, para obter estimativas de erro, utiliza-se em vez de $c = \|C\|$ e de $r_\sigma(C)$ uma outra constante c_r tal que,

$$r^{(k)} = \frac{\|x^{(k+1)} - x^{(k)}\|_V}{\|x^{(k)} - x^{(k-1)}\|_V} \leq c_r, \qquad \forall k > k_0,$$

quando for possível obtê-la experimentalmente.

Proposição. O método iterativo (#) converge para a solução x do sistema $Ax = b$, $\forall x^{(0)} \in \mathbb{C}^n$, se e só se todas as raízes $\lambda_1, \ldots, \lambda_n$ da equação polinomial

$$\det(\lambda M + N) = 0,$$

tiverem módulo inferior à unidade.

Dem.: $(\cdots)$

Proposição. Os métodos de Jacobi e Gauss-Seidel convergem para a solução do sistema $Ax = b$, $\forall x^{(0)} \in \mathbb{C}^n$, se e só se $r_\sigma(C_J) < 1$ e $r_\sigma(C_{GS}) < 1$, respectivamente.

Proposição. Os métodos de Jacobi e Gauss-Seidel convergem para a solução do sistema $Ax = b$, $\forall x^{(0)} \in \mathbb{C}^n$, se e só se todas as raízes das equações

$$\det(\lambda M_J + N_J) = 0, \qquad \det(\lambda M_{GS} + N_{GS}) = 0,$$

respectivamente, tiverem módulo inferior à unidade.

Definição.

(1) Diz-se que a matriz $A \in \mathbb{M}^n(\mathbb{C}), A = [a_{ij}]$ é uma **matriz de diagonal estritamente dominante por linhas** (MDEDL) se verificar as condições

$$|a_{ii}| > \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}|, \qquad \forall i \in \{1, 2, \ldots, n\}.$$

(2) Diz-se que a matriz $A \in \mathbb{M}^n(\mathbb{C}), A = [a_{ij}]$ é uma **matriz de diagonal estritamente dominante por colunas** (MDEDC) se verificar as condições

$$|a_{jj}| > \sum_{\substack{i=1 \\ i \neq j}}^{n} |a_{ij}|, \qquad \forall j \in \{1, 2, \ldots, n\}.$$

Proposição. Seja $A \in \mathbb{M}^n(\mathbb{C})$ uma MDEDL ou MDEDC. Então A é não singular.

Dem.: $(\cdots)$

Proposição. Seja $A \in \mathbb{M}^n(\mathbb{C})$ uma MDEDL. Definam-se as quantidades $\alpha_1, \ldots, \alpha_n$, $\beta_1, \ldots, \beta_n$ por

$$\alpha_1 = 0, \qquad \alpha_i = \sum_{j=1}^{i-1} \left| \frac{a_{ij}}{a_{ii}} \right|, \qquad i \in \{2, \ldots, n\},$$

$$\beta_i = \sum_{j=i+1}^{n} \left| \frac{a_{ij}}{a_{ii}} \right|, \qquad i \in \{1, \ldots, n-1\}, \qquad \beta_n = 0.$$

Então:

$$(1)\ \ \alpha_i + \beta_i < 1, \quad \forall i \in \{1, \ldots, n\}; \qquad (2)\ \ \frac{\beta_i}{1 - \alpha_i} < 1, \quad \forall i \in \{1, \ldots, n\}.$$

Dem.: $(\cdots)$

Proposição. Seja $A \in \mathbb{M}^n(\mathbb{C})$ uma MDEDL. Definam-se as quantidades

$$\mu = \max_{1 \leq i \leq n} (\alpha_i + \beta_i) < 1, \qquad \eta = \max_{1 \leq i \leq n} \frac{\beta_i}{1 - \alpha_i} < 1.$$

Então

$$(1)\ \ \|C_J\|_\infty = \mu; \qquad (2)\ \ \eta \leq \mu; \qquad (3)\ \ \|C_{GS}\|_\infty \leq \eta.$$

Dem.: (1) $(\cdots)$ (2) $(\cdots)$

Proposição. Seja $A \in \mathbb{M}^n(\mathbb{C})$ uma MDEDL. Então o método de Jacobi converge para a solução x do sistema $Ax = b$, $\forall x^{(0)} \in \mathbb{C}^n$, e é válida a estimativa de erro

$$\|e^{(k+1)}\|_\infty \leq \mu \|e^{(k)}\|_\infty.$$

Proposição. Seja $A \in \mathbb{M}^n(\mathbb{C})$ uma MDEDL. Então o método de Gauss-Seidel converge para a solução x do sistema $Ax = b$, $\forall x^{(0)} \in \mathbb{C}^n$, e é válida a estimativa de erro

$$\|e^{(k+1)}\|_\infty \leq \eta \|e^{(k)}\|_\infty.$$

Nota. Uma vez que $r_\sigma(C_J) \leq \|C_J\|_\infty$ e $r_\sigma(C_{GS}) \leq \|C_{GS}\|_\infty$ este resultado pode levar a pensar que sendo A uma MDEDL então também se verifica que $r_\sigma(C_{GS}) \leq r_\sigma(C_J) < 1$, isto é, que o método de Gauss-Seidel converge pelo menos tão rapidamente quanto o método de Jacobi. Isto não é verdade em geral, mas apenas impondo outras condições a A. Assim, por exemplo, é verdadeiro o seguinte resultado:

Proposição (Teorema de Stein-Rosenberg). Seja $A \in \mathbb{M}^n(\mathbb{C})$ uma matriz tal que a matriz iteradora do método de Jacobi é não negativa, $C_J \geq 0$ (isto é, $(C_J)_{ij} \geq 0, \forall i, j$). Seja C_{GS} a matriz iteradora do método de Gauss-Seidel. Então uma e uma só das seguintes proposições é verdadeira:

(1) $r_\sigma(C_J) = r_\sigma(C_{GS}) = 0$;

(2) $0 < r_\sigma(C_{GS}) < r_\sigma(C_J) < 1$;

(3) $r_\sigma(C_J) = r_\sigma(C_{GS}) = 1$;

(4) $1 < r_\sigma(C_J) < r_\sigma(C_{GS})$.

Nota. A condição $C_J \geq 0$ é satisfeita em particular se A é tal que $D_A > 0$ e $A - D_A \leq 0$. Uma vez que esta condição é verificada para quase todos os sistemas lineares que são obtidos pela aproximação às diferenças finitas de operadores diferenciais lineares, este teorema dá-nos em muitos casos práticos a informação importante de que, quando convergem, o método de Gauss-Seidel converge mais depressa do que o método de Jacobi.

Proposição. Seja $A \in \mathbb{M}^n(\mathbb{C})$ uma MDEDC. Então quer o método de Jacobi quer o método de Gauss-Seidel convergem para a solução x do sistema $Ax = b$, $\forall x^{(0)} \in \mathbb{C}^n$.

Definição. Uma matriz hermitiana $A \in \mathbb{M}^n(\mathbb{C})$ diz-se **definida positiva** se

$$x^* A x > 0, \qquad \forall x \in \mathbb{C}^n \setminus \{0\}.$$

Proposição. Cada uma das seguintes condições é necessária e suficiente para que uma matriz hermitiana $A \in \mathbb{M}^n(\mathbb{C})$ seja definida positiva:

(1) Todos os valores próprios de A são positivos.

(2) Todos os menores principais de A são positivos.

Nota. Chama-se submatriz principal de $A \in \mathbb{M}^n(\mathbb{C})$ à matriz $A_k \in \mathbb{M}^k(\mathbb{C})$, $k \in \{1, 2, \ldots, n\}$ cujos elementos são os elementos das primeiras k linhas e k colunas de A. Chama-se **menor principal** de A a $\det A_k$.

Proposição. Seja $A \in \mathbb{M}^n(\mathbb{C})$ uma matriz hermitiana e definida positiva. Então o método de Gauss-Seidel converge para a solução x do sistema $Ax = b$, $\forall x^{(0)} \in \mathbb{C}^n$.

Proposição. Seja $A \in \mathbb{M}^n(\mathbb{C})$ uma matriz hermitiana e definida positiva e tal que a matriz $2D_A - A$ é definida positiva. Então o método de Jacobi converge para a solução x do sistema $Ax = b$, $\forall x^{(0)} \in \mathbb{C}^n$.

• A análise de convergência do método SOR é mais complicada que a do método de Jacobi com relaxação pois a matriz iteradora depende não linearmente do parâmetro de relaxação:

$$C_{Jm}(\omega) = I - \omega D^{-1}A, \qquad C_{SOR}(\omega) = I - \omega(D + \omega L)^{-1}A.$$

Proposição. O método de Jacobi modificado é convergente se e só se todos os valores próprios da matriz C_J do método de Jacobi tiverem partes reais inferiores à unidade. A convergência verifica-se para $\omega \in]0, \omega_{\sup}[$, onde

$$\omega_{\sup} = \min_{\lambda \in \sigma(C_J)} \frac{2b(\lambda)}{2b(\lambda) + |\lambda|^2 - 1}, \qquad b(\lambda) = 1 - \Re(\lambda),$$

Proposição. Se o método de Jacobi for convergente então o método de Jacobi modificado com $\omega \in]0, 1]$ também é convergente.

Proposição (Teorema de Kahan). Seja $A \in \mathbb{M}^n(\mathbb{C})$. É condição necessária para que o método SOR convirja para a solução x do sistema $Ax = b$, $\forall x^{(0)} \in \mathbb{C}^n$, que $\omega \in]0, 2[$.

Nota. Prova-se que

$$r_\sigma(C_{SOR}(\omega)) \geq |\omega - 1|, \qquad \forall \omega \in \mathbb{R},$$

e portanto que

$$r_\sigma(C_{SOR}(\omega)) \geq 1, \qquad \forall \omega \notin]0, 2[.$$

Proposição (Teorema de Ostrowski-Reich). Seja $A \in \mathbb{M}^n(\mathbb{C})$ uma matriz hermitiana e definida positiva. É condição suficiente para que o método SOR convirja para a solução x do sistema $Ax = b$, $\forall x^{(0)} \in \mathbb{C}^n$, que $\omega \in]0, 2[$.

Nota. Prova-se que

$$r_\sigma(C_{SOR}(\omega)) < 1, \qquad \forall \omega \in]0, 2[.$$

Nota. Em particular este resultado significa que o método de Gauss-Seidel, obtido para $\omega = 1$, converge para matrizes hermiteanas e definidas positivas.

• O cálculo do parâmetro de relaxação óptimo, isto é, o parâmetro que minimiza o raio espectral, é difícil excepto nalguns casos particulares. Geralmente obtém-se apenas um valor aproximado experimentando numericamente diversos valores de ω e observando o

efeito na rapidez de convergência. Este esforço é justificado pois o aumento da rapidez de convergência pode ser significativo. Damos seguidamente um exemplo de uma classe de matrizes para a qual o problema pode ser tratado analiticamente com facilidade. Um outro caso que também pode ser tratado analiticamente é o de uma outra classe de matrizes que ocorre na discretização de problemas de valor na fronteira, as chamadas *matrizes consistentemente ordenadas*.

Proposição. Suponhamos que a matriz iteradora C_J do método de Jacobi tem valores próprios reais e inferiores à unidade e designemos por λ_1 e λ_n o maior e o menor destes valores, respectivamente. Então o método de Jacobi modificado com $\omega \in]0, \omega_{\text{sup}}[$, onde $\omega_{\text{sup}} = \dfrac{2}{1-\lambda_n}$, é convergente. Além disso o raio espectral de $C_{Jm}(\omega)$ tem o valor mínimo para $\omega_{\text{opt}} = \dfrac{2}{2-\lambda_1-\lambda_n}$ e esse valor mínimo é

$$r_\sigma(C_{Jm}(\omega_{\text{opt}})) = \frac{\lambda_1 - \lambda_n}{2-\lambda_1-\lambda_n}.$$

Dem.: $(\cdots)$

Exemplo. Considere-se o sistema linear $Ax = b$, onde

$$A = \begin{bmatrix} 10 & 3 & 1 \\ 2 & -10 & 3 \\ 1 & 3 & 10 \end{bmatrix}, \qquad b = \begin{bmatrix} 14 \\ -5 \\ 14 \end{bmatrix}.$$

(a) Determinar as seis primeiras iteradas do método de Jacobi com condição inicial $x^{(0)} = [0 \;\; 0 \;\; 0]^T$. Justificar a convergência do método e obter uma estimativa do erro da iterada $x^{(6)}$.

(b) Idem, para o método de Gauss-Seidel.

Resolução:

(a) $x^{(k+1)} = D^{-1}[b - (L+U)x^{(k)}]$

$$x^{(k+1)} = \frac{1}{10}\begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix}\left(\begin{bmatrix} 14 \\ -5 \\ 14 \end{bmatrix} - \begin{bmatrix} 0 & 3 & 1 \\ 2 & 0 & 3 \\ 1 & 3 & 0 \end{bmatrix} x^{(k)}\right)$$

$$\begin{cases} x_1^{(k+1)} = \dfrac{1}{10}\left(14 - 3x_2^{(k)} - x_3^{(k)}\right) \\ x_2^{(k+1)} = \dfrac{1}{10}\left(5 + 2x_1^{(k)} + 3x_3^{(k)}\right) \\ x_3^{(k+1)} = \dfrac{1}{10}\left(14 - x_1^{(k)} - 3x_2^{(k)}\right) \end{cases}$$

k	$x_1^{(k)}$	$x_2^{(k)}$	$x_3^{(k)}$
0	0.000000000	0.000000000	0.000000000
1	1.400000000	0.500000000	1.400000000
2	1.110000000	1.200000000	1.110000000
3	0.929000000	1.055000000	0.929000000
4	0.990600000	0.964500000	0.990600000
5	1.011590000	0.995300000	1.011590000
6	1.000251000	1.005795000	1.000251000

(b) $x^{(k+1)} = D^{-1}[b - Lx^{(k+1)} - Ux^{(k)}]$

$$x^{(k+1)} = \frac{1}{10}\begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix}\left(\begin{bmatrix} 14 \\ -5 \\ 14 \end{bmatrix} - \begin{bmatrix} 0 & 0 & 0 \\ 2 & 0 & 0 \\ 1 & 3 & 0 \end{bmatrix} x^{(k+1)} - \begin{bmatrix} 0 & 3 & 1 \\ 0 & 0 & 3 \\ 0 & 0 & 0 \end{bmatrix} x^{(k)}\right)$$

$$\begin{cases} x_1^{(k+1)} = \frac{1}{10}\left(14 - 3x_2^{(k)} - x_3^{(k)}\right) \\ x_2^{(k+1)} = \frac{1}{10}\left(5 + 2x_1^{(k+1)} + 3x_3^{(k)}\right) \\ x_3^{(k+1)} = \frac{1}{10}\left(14 - x_1^{(k+1)} - 3x_2^{(k+1)}\right) \end{cases}$$

k	$x_1^{(k)}$	$x_2^{(k)}$	$x_3^{(k)}$
0	0.000000000	0.000000000	0.000000000
1	1.400000000	0.780000000	1.026000000
2	1.063400000	1.020480000	0.987516000
3	0.995104400	0.995275680	1.001906856
4	1.001226610	1.000817379	0.999632125
5	0.999791574	0.999847952	1.000066457
6	1.000038969	1.000027731	0.999987784

Convergência: os métodos de Jacobi e Gauss-Seidel são convergentes para a solução do sistema para qualquer condição inicial pois A é MDEDL.

Estimativas de erro:

$$\|e^{(k)}\|_\infty \leq \frac{c}{1-c}\|x^{(k)} - x^{(k-1)}\|_\infty$$

$c = \mu = \max_{1\leq i\leq n}(\alpha_i + \beta_i)$ Método de Jacobi

$c = \eta = \max_{1\leq i\leq n}\frac{\beta_i}{1-\alpha_i}$ Método de Gauss-Seidel

i	α_i	β_i	$\alpha_i + \beta_i$	$\dfrac{\beta_i}{1-\alpha_i}$
1	0	$\frac{4}{10}$	$\frac{4}{10}$	$\frac{4}{10}$
2	$\frac{2}{10}$	$\frac{3}{10}$	$\frac{5}{10}$	$\frac{3}{8}$
3	$\frac{4}{10}$	0	$\frac{4}{10}$	0

$$\mu = \frac{1}{2}, \qquad \eta = \frac{2}{5}$$

Método de Jacobi:

$$x^{(6)} - x^{(5)} = \begin{bmatrix} -0.0113390 \\ 0.0104950 \\ -0.0113390 \end{bmatrix}, \qquad \|x^{(6)} - x^{(5)}\|_\infty = 0.0113390$$

$$\|e^{(6)}\|_\infty \leq \|x^{(6)} - x^{(5)}\|_\infty = 0.0113390$$

Com $c = r_\sigma(C_J) = \dfrac{\sqrt{15}}{10} = 0.387298$ obtém-se

$$\|e^{(6)}\|_\infty \leq 0.00716755$$

Estes valores devem ser comparado entre si e com o "erro verdadeiro"

$$\|e^{(6)}\|_\infty = 0.005795$$

Método de Gauss-Seidel:

$$x^{(6)} - x^{(5)} = \begin{bmatrix} 0.000247395 \\ 0.000179779 \\ -0.000078673 \end{bmatrix}, \qquad \|x^{(6)} - x^{(5)}\|_\infty = 0.000247395$$

$$\|e^{(6)}\|_\infty \leq \frac{2}{3}\|x^{(6)} - x^{(5)}\|_\infty = 0.000164930$$

Com $c = r_\sigma(C_{GS}) = \dfrac{\sqrt{67 + \sqrt{13489}}}{1000} = 0.183142$ obtém-se

$$\|e^{(6)}\|_\infty \leq 0.0000554667$$

Estes valores devem ser comparado entre si e com o "erro verdadeiro"

$$\|e^{(6)}\|_\infty = 0.000038969$$

Exemplo. Considere-se o sistema linear $Ax = b$, onde

$$A = \begin{bmatrix} 2 & 1 & 0 \\ -1 & 2 & 1 \\ 0 & -1 & 2 \end{bmatrix}, \qquad b = \begin{bmatrix} 2 \\ 2 \\ 1 \end{bmatrix}$$

(a) Determinar a solução aproximada $x^{(k)}$ do sistema pelo método de Jacobi com condição inicial $x^{(0)} = [0.5 \;\; 0.8 \;\; 1.0]^T$ tal que é satisfeita a desigualdade

$$\|x^{(k)} - x^{(k-1)}\|_2 \leq 0.01.$$

Obter uma estimativa do erro da iterada $x^{(k)}$.

(b) Idem, para o método de Gauss-Seidel.

Resolução:

(a) $x^{(k+1)} = D^{-1}[b - (L+U)x^{(k)}]$

$$x^{(k+1)} = \frac{1}{2}\left(\begin{bmatrix} 2 \\ 2 \\ 1 \end{bmatrix} - \begin{bmatrix} 0 & 1 & 0 \\ -1 & 0 & 1 \\ 0 & -1 & 0 \end{bmatrix} x^{(k)}\right)$$

$$\begin{cases} x_1^{(k+1)} = \frac{1}{2}\left(2 - x_2^{(k)}\right) \\ x_2^{(k+1)} = \frac{1}{2}\left(2 + x_1^{(k)} - x_3^{(k)}\right) \\ x_3^{(k+1)} = \frac{1}{2}\left(1 + x_2^{(k)}\right) \end{cases}$$

k	$x_1^{(k)}$	$x_2^{(k)}$	$x_3^{(k)}$	$\|x^{(k)} - x^{(k-1)}\|_2$	$\frac{\|x^{(k)} - x^{(k-1)}\|_2}{\|x^{(k-1)} - x^{(k-2)}\|_2}$
0	0.5	0.8	1.0		
1	0.6	0.75	0.9	0.15	
2	0.625	0.85	0.875	0.106066	0.707107
3	0.575	0.875	0.925	0.07500	0.707107
4	0.5625	0.825	0.9375	0.053033	0.707107
5	0.5875	0.8125	0.9125	0.03750	0.707107
6	0.59375	0.8375	0.90625	0.0265165	0.707107
7	0.58125	0.84375	0.91875	0.01875	0.707107
8	0.578125	0.83125	0.921875	0.0132583	0.707107
9	0.584375	0.828125	0.915625	0.009375	0.707107

$$\|e^{(9)}\|_2 \leq \frac{c}{1-c}\|x^{(9)} - x^{(8)}\|_2$$

$$c = r_\sigma(C_J) = \frac{1}{\sqrt{2}}$$

$$\|e^{(9)}\|_2 \leq 0.0242$$

(b) $x^{(k+1)} = D^{-1}[b - Lx^{(k+1)} - Ux^{(k)}]$

$$x^{(k+1)} = \frac{1}{2}\left(\begin{bmatrix} 2 \\ 2 \\ 1 \end{bmatrix} - \begin{bmatrix} 0 & 0 & 0 \\ -1 & 0 & 0 \\ 0 & -1 & 0 \end{bmatrix} x^{(k+1)} - \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} x^{(k)}\right)$$

$$\begin{cases} x_1^{(k+1)} = \frac{1}{2}\left(2 - x_2^{(k)}\right) \\ x_2^{(k+1)} = \frac{1}{2}\left(2 + x_1^{(k+1)} - x_3^{(k)}\right) \\ x_3^{(k+1)} = \frac{1}{2}\left(1 + x_2^{(k+1)}\right) \end{cases}$$

k	$x_1^{(k)}$	$x_2^{(k)}$	$x_3^{(k)}$	$\|x^{(k)} - x^{(k-1)}\|_2$	$\frac{\|x^{(k)} - x^{(k-1)}\|_2}{\|x^{(k-1)} - x^{(k-2)}\|_2}$
0	0.5	0.8	1.0		
1	0.6	0.8	0.9	0.141421	
2	0.6	0.85	0.925	0.0559017	0.395285
3	0.575	0.825	0.9125	0.037500	0.67082
4	0.5875	0.8375	0.91875	0.018750	0.5
5	0.58125	0.83125	0.915625	0.009375	0.5

$$\|e^{(5)}\|_2 \leq \frac{c}{1-c}\|x^{(5)} - x^{(4)}\|_2$$

$$c = r_\sigma(C_{GS}) = \frac{1}{2}$$

$$\|e^{(5)}\|_2 \leq 0.01$$

Exemplo. Considere-se o sistema linear $Ax = b$, onde

$$A = \begin{bmatrix} 2 & 1 & 1 \\ 1 & 3 & 1 \\ 1 & 2 & 2 \end{bmatrix}.$$

(a) Determinar o valor $\omega = \omega_{\text{opt}}$ para o qual o método de Jacobi com relaxação converge mais rapidamente e o valor de $r_\sigma(C_{Jm}(\omega_{\text{opt}}))$.

(b) Idem, para o método de Gauss-Seidel com relaxação.

Resolução:

(a)

$$C_{Jm}(\omega) = I - \omega D^{-1}A = \begin{bmatrix} 1-\omega & -\frac{\omega}{2} & -\frac{\omega}{2} \\ -\frac{\omega}{2} & 1-\omega & -\frac{\omega}{3} \\ -\frac{\omega}{2} & -\omega & 1-\omega \end{bmatrix}$$

$$\sigma\left(C_{Jm}(\omega)\right) = \left\{1 - \frac{\omega}{2}, 1 - \frac{\omega}{2}, 1 - 2\omega\right\}$$

$$r_\sigma\left(C_{Jm}(\omega)\right) = \begin{cases} 1 - 2\omega, & \omega \leq 0, \\ 1 - \frac{\omega}{2}, & 0 \leq \omega \leq \omega_{\text{opt}}, \\ 2\omega - 1, & \omega_{\text{opt}} \leq \omega \end{cases}$$

$\omega_{\text{opt}} = \dfrac{4}{5}, \qquad r_\sigma\left(C_{Jm}(\omega_{\text{opt}})\right) = \dfrac{3}{5}$

O método converge para $\omega \in]0,1[$, intervalo em que $r_\sigma\left(C_{Jm}(\omega)\right) < 1$.

(b)

$C_{\text{SOR}}(\omega) = I - \omega(D + \omega L)^{-1}A$

$$= \begin{bmatrix} 1-\omega & -\frac{\omega}{2} & -\frac{\omega}{2} \\ -\frac{\omega}{3}(1-\omega) & \frac{1}{6}(6-6\omega+\omega^2) & -\frac{\omega}{6}(2-\omega) \\ -\frac{\omega}{6}(3-5\omega+2\omega^2) & -\frac{\omega}{12}(12-15\omega+2\omega^2) & \frac{1}{12}(12-12\omega+7\omega^2-2\omega^3) \end{bmatrix}$$

$$\begin{aligned} \det(C_{\text{SOR}}(\omega) - \lambda I) = \; & (1-\omega)^3 + \frac{1}{12}(-36+72\omega-45\omega^2+8\omega^3)\lambda \\ & + \frac{1}{12}(36-36\omega+9\omega^2-2\omega^3)\lambda^2 - \lambda^3 \end{aligned}$$

$r_\sigma(C_{\text{SOR}}(\omega)) = \max_{i=1,2,3}\{|\lambda_i(\omega)\}$

$\omega_{\text{opt}} = \{\omega \in \mathbb{R}^+ : r_\sigma(C_{\text{SOR}}(\omega)) \text{ é mínimo }\}$

ω	$r_\sigma(C_{\text{SOR}}(\omega))$
0.	1.
0.25	0.880321
0.5	0.75
0.75	0.591443
1.	0.333333
1.05	0.252151
1.1	0.239192
1.15	0.252158
1.2	0.279534
1.25	0.313121
1.5	0.5
1.75	0.9061
1.8	0.994896
1.9	1.18037
2.0	1.3767

$\omega_{\text{opt}} = 1.1, \qquad r_\sigma(C_{\text{SOR}}(\omega_{\text{opt}})) = 0.239192$

O método é convergente para $\omega \in]0, 1.8]$.

Determinação experimental do valor óptimo de ω para ambos os métodos:

$$Ax = b, \qquad b = \begin{bmatrix} 4 \\ 5 \\ 5 \end{bmatrix}, \qquad x^{(0)} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

ω	$k < 200 : \|x^{(k)} - x^{(k-1)}\|_2 \leq 10^{-5}$	
	Método Jm	Método SOR
0.1	175	169
0.2	94	86
0.3	64	56
0.4	49	38
0.5	39	26
0.6	33	25
0.7	28	22
0.8	**26**	19
0.9	58	16
1.0		13
1.1		**11**
1.2		12
1.3		14
1.4		17
1.5		19
1.6		31
1.7		64

Comparação entre os métodos directos e os métodos iterativos

Notas.

(a) Os métodos directos requerem $\approx \frac{n^3}{3}$ operações para obter a solução. Os métodos iterativos implicam normalmente um cálculo de $\approx n^2$ operações em cada iteração, o que os torna ineficazes face aos métodos directos para $n_{\text{it}} > \frac{n}{3}$. Por outro lado, a precisão atingida ao fim de $\frac{n}{3}$ iteradas não é normalmente muito boa ($\|x - x^{(n/3)}\|_V \leq c^{n/3}\|x - x^{(0)}\|_V$). Por estas razões os métodos iterativos só se tornam realmente eficazes para matrizes de grandes dimensões e, em especial, quando as matrizes são esparsas (isto é, matrizes com poucos elementos diferentes de zero). Nestes casos os métodos directos não se simplificam em geral muito enquanto que os métodos iterativos apresentam uma redução apreciável do número de operações.

(b) Os métodos iterativos para sistemas lineares convergentes são *estáveis*, isto é, partindo de dois vectores iniciais próximos, $\xi^{(0)}$ e $\eta^{(0)}$, obtêm-se sempre duas sucessões $\{x^{(k)}\}$ e $\{y^{(k)}\}$ igualmente próximas, convergindo ambas para o mesmo vector x, solução exacta do sistema linear, verificando-se:

$$\exists \theta > 0 : \max_k \|x^{(k)} - y^{(k)}\| \leq \theta \|\xi^{(0)} - \eta^{(0)}\|, \quad \forall \xi^{(0)}, \eta^{(0)} \in \mathbb{C}^n.$$

Na presença de erros de arredondamento os métodos iterativos, desde que aplicados a sistemas bem condicionados, continuam estáveis. Isto significa que não há perigo de os erros de arredondamento cometidos no cálculo poderem resultar em erros significativos no resultado final. Isto representa uma importante vantagem dos métodos iterativos sobre os métodos directos em que os erros de arredondamento

podem propagar-se ao longo do cálculo, conduzindo a erros muito grandes no resultado final, mesmo que os sistemas sejam bem condicionados, e sobretudo quando se tratam de sistemas de grandes dimensões.